Name : Samiksha Sudhir Murkute Roll no: ET1-50 PRN: 202401070099 problem statement: CRICKET DATA ANALYSIS

```
from google.colab import files
uploaded = files.upload()
```

Choose Files world_cup_data.csv

- **world_cup_data.csv**(text/csv) - 12777 bytes, last modified: 4/28/2025 - 100% done

```
Saving world_cup_data.csv to world_cup_data (2).csv
```

1.Identify the top 5 players who have scored the most runs in all Cricket World Cup matches combined.

```
import pandas as pd
df = pd.read_csv('world_cup_data.csv')
top_scorers = df.groupby('player')['runs'].sum().sort_values(ascending=False).head(5)
print(top_scorers)
```

```
player
Virat Kohli        1894
Steve Smith        1849
Kane Williamson    1723
Babar Azam         1515
Joe Root           1417
Name: runs, dtype: int64
```

2.Find the top 5 bowlers who have taken the most wickets in the tournament history.

```
top_bowlers = df.groupby('player')['wickets'].sum().sort_values(ascending=False).head(5)
print(top_bowlers)
```

```
player
Virat Kohli        58
Steve Smith        55
Kane Williamson    43
Babar Azam         34
Joe Root           25
Name: wickets, dtype: int64
```

3.Calculate how many unique matches each team has played.

```
total_matches = df.groupby('team')['match_id'].nunique()
print(total_matches)
```

```
team
Australia      13
England         7
India          14
New Zealand    15
Pakistan       10
Name: match_id, dtype: int64
```

4.Determine the total number of runs scored by each team across all matches.

```
team_runs = df.groupby('team')['runs'].sum().sort_values(ascending=False)
print(team_runs)
```

```
team
New Zealand    2471
India          1962
Australia      1825
Pakistan       1432
England         708
Name: runs, dtype: int64
```

5.Find the average runs scored per match for each player.

```
avg_runs = df.groupby('player')['runs'].mean().sort_values(ascending=False)
print(avg_runs)
```

```
player
Joe Root           101.214286
Babar Azam          94.687500
Steve Smith         80.391304
Kane Williamson     78.318182
Virat Kohli         75.760000
Name: runs, dtype: float64
```

6.Identify the top 5 players who have hit the maximum number of sixes.

```
most_sixes = df.groupby('player')['sixes'].sum().sort_values(ascending=False).head(5)
print(most_sixes)
```

```
player
Steve Smith        49
Kane Williamson    41
Virat Kohli        36
Joe Root           34
Babar Azam         31
Name: sixes, dtype: int64
```

7.Calculate the average batting strike rate for each player.

```
df['strike_rate'] = np.where(df['balls_faced'] > 0, (df['runs'] / df['balls_faced']) * 100, 0)
player_sr = df.groupby('player')['strike_rate'].mean().sort_values(ascending=False).head(10)
print(player_sr)
```

```
player
Joe Root           1099.051210
Steve Smith         507.945076
Babar Azam          459.822128
Virat Kohli         408.670113
Kane Williamson     153.561712
Name: strike_rate, dtype: float64
```

8.Determine how many matches were played in each World Cup year.

```
df['year'] = pd.to_datetime(df['date']).dt.year
matches_per_year = df.groupby('year')['match_id'].nunique()
print(matches_per_year)
```

```
year
2019    20
Name: match_id, dtype: int64
```

9.Find the most frequent mode of dismissal (e.g., bowled, caught, run-out, etc.) in the dataset.

```
dismissal_count = df['dismissal_type'].value_counts()
print(dismissal_count)
```

```
dismissal_type
bowled     21
stumped    18
not out    17
caught     17
run out    14
lbw        13
Name: count, dtype: int64
```

10.List the top 5 players who have scored zero (duck) most often

```
ducks = df[df['runs'] == 0].groupby('player').size().sort_values(ascending=False).head(5)
print(ducks)
```

```
player
Babar Azam     1
Virat Kohli    1
```

```
dtype: int64
```

11.Find the bowlers with the best (lowest) average economy rate.

```
df['economy'] = df['runs_conceded'] / df['overs_bowled']
economy_avg = df.groupby('player')['economy'].mean().sort_values().head(5)
print(economy_avg)
```

```
player
Kane Williamson    14.817456
Steve Smith        17.736591
Virat Kohli        26.619073
Joe Root           49.326309
Babar Azam         62.226845
Name: economy, dtype: float64
```

12.Determine which players have received the most "Player of the Match" awards.

```
pom_awards = df['player_of_match'].value_counts().head(10)
print(pom_awards)
```

```
player_of_match
Steve Smith        23
Kane Williamson    23
Virat Kohli        21
Babar Azam         20
Joe Root           13
Name: count, dtype: int64
```

13.Identify the team with the most number of wins in the World Cup history.

```
wins = df[df['result'] == 'win'].groupby('team').size().sort_values(ascending=False)
print(wins)
```

```
team
New Zealand    16
Australia      15
India          12
Pakistan       11
England         4
dtype: int64
```

14.Find the top 5 venues that have hosted the maximum number of matches.

```
venue_count = df['venue'].value_counts().head(5)
print(venue_count)
```

```
venue
Lord's          22
MCG             22
SCG             21
Old Trafford    19
Eden Gardens    16
Name: count, dtype: int64
```

15.Analyze how each team's total runs have varied across different World Cup editions.

```
team_yearly_runs = df.groupby(['year', 'team'])['runs'].sum().unstack().fillna(0)
print(team_yearly_runs)
```

```
team  Australia  England  India  New Zealand  Pakistan
year
2019       1825      708   1962         2471      1432
```

16.Create a histogram showing the distribution of runs scored using NumPy.

```
import numpy as np
run_array = df['runs'].values
run_hist, bins = np.histogram(run_array, bins=[0, 10, 20, 30, 50, 100, 200])
print(run_hist)
```

```
[ 8  6  4  9 29 44]
```

17.Find the top 5 players who have appeared in the most number of matches.

```
most_matches = df.groupby('player')['match_id'].nunique().sort_values(ascending=False).head(5)
print(most_matches)
```

```
player
Virat Kohli        16
Steve Smith        14
Kane Williamson    14
Babar Azam         12
Joe Root            8
Name: match_id, dtype: int64
```

18.Identify the match where the highest combined number of runs was scored.

```

```

```
highest_run_match = df.groupby('match_id')['runs'].sum().sort_values(ascending=False).head(1)
print(highest_run_match)
```

```
match_id
1014    904
Name: runs, dtype: int64
```

19.Compare the head-to-head win record between two specific teams (e.g., India vs Australia).

```
team1, team2 = 'India', 'Australia'
head_to_head = df[((df['team'] == team1) & (df['opponent'] == team2)) |
                  ((df['team'] == team2) & (df['opponent'] == team1))]
head_to_head_result = head_to_head['winner'].value_counts()
print(head_to_head_result)
```

```
winner
Pakistan       6
England        4
New Zealand    3
India          2
Australia      1
Name: count, dtype: int64
```

20.Find out how many matches were decided via a Super Over.

```
super_over_matches = df[df['super_over'] == True]['match_id'].nunique()
print(super_over_matches)
```

```
19
```